報道元と報道時期を利用したWeb記事による将来予測

加藤　郁之† 　吉川　正俊†† 　加藤　誠†††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
††† 京都大学国際高等教育院／ JST さきがけ　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †fumiyuki@db.soc.i.kyoto-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp, †††kato@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，Web 上の記事を用いて将来における知識ベースの変化を予測する問題に取り組む．特に，記事の報道元と報道時期が，記事の内容から予測される結果の確信度に大きな影響を与えると仮定し，この仮定を明示的に表現する損失関数および特徴量を提案する．実験では,「サッカー選手の移籍予測」を題材として，記事の報道元と報道時期が予測精度に与える影響，および，提案する損失関数によって結果の確信度が正確に予測できるかを評価した．

キーワード　Web 記事，将来予測，テキストマイニング

1 はじめに

何らかの意思決定を行う際には，現在の状況だけでなく，将来どのような変化が起こりうるのかを予測することが重要となる．将来の予測を行うためには現状を網羅的に把握する必要があるが，Web 上で発信されるニュース記事はこれを行うためのもっとも有用なソースの 1 つである．例えば，Web である企業の買収に関するトピックを検索して Web 上の記事 (以下 Web 記事) を読み，その企業が買収される可能性が高いかを予測し，その企業へ投資を行うべきかどうかを判断する．

しかしながら，Web 記事から将来を予測することは容易ではなく，以下のような問題点が挙げられる．1 つは，Web から得られる記事が非常に膨大である点である．全ての記事をを精読し，それらから総合的に予測を行うのは容易ではない．もう 1 つは，同一トピックであっても，各記事ごとに主張やその根拠，その記事に書かれている内容の確からしさ，確信度などが異なっている点である．これは，報道メディア一般に言えることだが，Web において特に顕著である．同一トピックに対する異なる見解を持った記事が多数存在すれば情報の取捨選択が必要となる．記事の多様性を理解するためには，記事のコンテクストを考慮した上でその情報の真偽を判断する必要があり，正確な意思決定の難易度は上がる．

そこで本研究では，機械学習によってニュース記事に基づいた将来予測を行うことを目的とする．特に，事実の変化を予測することに焦点を当て，ある時刻における知識ベースが別の時刻においてどのように変化するかを予測するタスクに取り組む．このタスクにおける入力は Web 記事の集合であり，出力は将来の知識ベースに追加されるであろう新規の「知識」である．

本研究おいては，Web 記事を利用する際にそのテキストデータを入力とすることに加えて，記事の報道元や報道時期を利用することで，その記事の確信度を学習し将来予測を行う．具体的には，記事の報道元と報道時期が，記事の内容から予測される結果の確信度に大きな影響を与えると仮定し，この仮定を明示的に表現する損失関数および特徴量を提案する．

実験では，サッカー選手の移籍を題材にして将来予測を行う．この題材の特徴としては，主に各報道元によって記事が多様であること，また，移籍にはあらかじめ定められた移籍可能期間があり，移籍が決定する直前に報道される記事の方が確信度が高いため，報道時期によって報道の確信度が顕著に異なることが挙げられる．Web から 20 の Web サイトを選定し，それぞれを報道元としてみなしてサッカー選手の移籍に関する Web 記事を収集した．我々は，Web 記事のテキストデータのみによって学習されたベースライン手法と，先に述べた提案手法を比較を行なった．実験の結果，Web 記事からの将来予測がある程度の精度で達成できること，テキストデータに加え，報道元と報道時期から作成した特徴量を加えることで予測精度が上がることが示唆された．提案損失関数を使用した場合には，その損失関数の収束を実験的に確認することができたとともに，ベースライン手法と比較して，それ以上の精度で予測を行えることを示した．

この論文における貢献を以下に示す : (1) 報道元と報道時期の情報を考慮して確信度を学習することで予測精度の向上を提案した．(2) 確信度を明示的に学習する損失関数を提案した．(3) 実際にサッカー選手の移籍に関する記事データとその正解データの収集を行なって実験を行なった．

本論文の構成は以下の通りである．2 節では，企業の買収予測を題材にした，Web 記事を利用した将来予測に関する関連研究，また，Web 記事を利用した掲載ドメインの信頼度を学習する研究について簡単に述べる．3 節では，本研究における問題設定，その解決策と提案手法について，4 節では本研究にて行う実験について，5 節では本論文のまとめについて述べる．

2 関連研究

本節では，Web 上のリソースから将来予測を行うことに関

する関連研究について述べる．本研究のようにサッカー選手の移籍予測を行う研究は散見されないが，異なる特定の対象に対し，Web 記事のコーパス等を利用してテキストデータを元に将来予測を行う一連の流れをもつ研究は多数見られる．

以下，Web 記事を会社の買収予測に利用した研究 [14] について述べるとともに，サッカー選手の移籍予測についても類似する性質を考える．会社の買収予測は，従来から研究が活発に行われきた領域である．予測の方法として主流であるものは，財政に関するな数値，経営に関する数値を数理モデルの入力とし，機械学習によってパラメータを推定する方法である．この入力に加えて，Web 記事を用いて新たな特徴量を作成したのがこの研究の特徴であり，特定の企業に関する記事集合から LDA [2] を用いてトピック抽出を行う．学習したモデルを利用して，各企業に関する Web 記事からトピックに関する特徴量を得る．その特徴量を，従来の特徴量に付加してベイジアンネットワークを用いて学習を行なう．その結果，特に必要な特徴量が欠けているデータに対して，従来の手法の結果と比べて精度が改善されることが示された．LDA により抽出されたトピックごとの，頻出上位の単語を元に，ソーシャルネットワーク，スタートアップ，広告事業などのトピックが抽出されていると解釈されている．この結果から，会社の傾向を Web 記事を通じてカテゴライズできたということが分かる．また，実験結果から，Web 記事から会社の傾向を掴んで，その情報が会社の買収予測の改善に寄与していると言える．Web 記事を利用した将来予測が可能であることを示唆した研究の 1 つである．この手法においては，記事のテキストデータをトピック抽出に用いているのに対して，本研究ではテキストデータを TF-IDF でベクトル化して特徴量としている．

企業の買収予測とサッカー選手の移籍予測は類似している部分がある．どちらもあるタイミングで明確に起こり，それまでは何も起こらない．形式的に言うと，状態が時系列で変遷していくが，状態が 2 状態しかなく，あるタイミングで 1 度しか変化しない．Web 記事はそれらの時系列の進度に追随して公開されるものであるために，これらの状態遷移予測に有効である．

次に，Web 記事の主張からその掲載元ドメインの信頼度を学習する手法 [13] について簡単に述べる．この手法では，様々な主張を掲載する Web ページを末端の子ノードとして，多数の Web ページを持つドメインに向けてのグラフを構成（論文中では Credibility Assessment グラフと呼ばれる）する．末端の各 Web ページの主張内容を評価して，その信頼性を統合的に上流のドメインに伝播させていくことで，ドメインの信頼性を学習する．複数のページから得られた信頼度を，一階述語論理を [0,1] の実数に緩めた Probabilistic Soft Logic (PSL) [6] によって統合し，ノード間で伝播させている．一般に，同一の Web サイト内においても，ある主張に対するいくつかのエビデンスが存在する．また，ドメインは当然ながら複数の Web ページを持っている．これらの関係を有向グラフに落とし込むことで情報源であるドメインを Web ページの文書の主張内容から遡って評価している．本研究では今回は 1 記事ごとにドメインの確信度の学習を行なったが，この研究内の手法のように，複数の記事からまとめて評価を行う方がより精度の高い評価が期待できる．ここで述べておきたいことは，この論文内において Web 記事の主張結果を評価することで，その元のドメインの信頼度を学習することができることが示唆された点である．この手法とは異なり，本研究においては，各記事のメタデータから記事の確信度に寄与すると期待される特量量作成し，損失関数を工夫することで確信度の学習を行う．

この他にも，いくつかの類似の研究がある．病気の蔓延，死者の増加，暴動の増加などの世界に起こる大きなイベントとそれらに関するニュース記事が類似の確率分布から生成されると仮定して，22 年分のニュースアーカイブと Web 記事を利用して，そのようなイベントが起こるかどうかの未来予測を行った研究 [12] や，ニュース記事情報とその報道のタイミングや報道場所から未来の紛争の予測や，逃亡者の居場所の予測などができることが示唆した研究 [8] が行われている．将来予測のタスクの対象は多岐に渡り，その手法やデータもいくつかの種類がある．映画の興行収入や成功を対象に予測を行なった研究では，近年多くの解析が行われている Twitter 等のソーシャルメディアを利用したもの [1]，映画に関する Web 上の大量のブログの感情分析結果を基にしたもの [11]，批評家達のレビューのテキストデータとそのメタデータを線形回帰の特徴量として予測を行なった，手法として本研究に非常に近い研究 [5] などがある．これらのトピックが Web の発展によって現在まで盛んに行われるようになったことがわかる．

3 手　　法

本節では，まず予測タスクの問題設定について述べ．次に，具体的なデータと問題の解決方法について説明を行う．

3.1 問 題 設 定

本研究では，将来予測に結果の有無やタイミングの判定を特定の知識ベースの知識の変化に対応させることで明確化する．知識ベース内の知識は下記の RDF トリプルモデル [7] によって表現される．

$$(s, p, o)$$

それぞれ s は主語，o は目的語，p は述語に対応している．これを用いて，問題設定を以下のように書くことができる．

まずは以下の変数を定義する．

- p' : 予測対象の述語
- KB_t : 時刻 t における知識ベース，

各 s に対する (s, p', o) を 1 つのみもつように制限

- $s \in S$: S 全ての主語の集合
- $o \in O$: O 全ての目的語の集合
- $t \in T$: T は全報道期間
- $a_{s,o}^{(t)} \in A_{s,o}^{(t)}$: $A_{s,o}^{(t)}$は記事集合

- $a_{s,o}^{(t)}$: 時刻 t における (s, p', o) に関連する記事
- τ : $(t - c \leqq \tau \leqq t)$,$t$ より前の一定の幅を持った期間

これらを用いて以下のように入力と出力を設定する．

入力　ある時刻 t,ある主語 s, s に対する記事集合 $A_{s,O}^{(\tau)} = (A_{s,o_1}^{(t_1)}, ..., A_{s,o_N}^{(t_M)})$
$o_1, ..., o_N$は当記事集合に現れた $o \in O$
$t_1, ..., t_M$は当記事集合に現れた報道時刻

出力　σ後の知識ベース $\mathrm{KB}_{t+\sigma}$がもつ (s, p', o) に含まれる o

このように 問題を知識ベースによって定式化することができ，正しい o を出力することを目標とする．本研究において具体的には主語 s がサッカー選手に対応し，目的語 $o \in O$ がチームに対応し，述語 p は「所属」で固定であり (s, p', o) が s が o に所属しているという事実に対応する．よって，選手 s に対して，時刻 t における各チームに関する記事集合 $A_{s,O}^{(\tau)}$ を元に，一定時間後にどのチームに所属しているのかを予測する問題であると定義できる．ただし，上で定義した通り，$(t - c \leqq \tau \leqq t)$ を満たす．この問題設定の場合は選手がどこへも移籍しない場合は，元々選手が所属していたチームが出力すべきチームになる．以下に例をあげると，当時プロサッカーチームである FC バルセロナに所属するリオネル・メッシ選手に対して，2015 年 7 月 1 日-2015 年 8 月 31 日の移籍期間に対応する 2015 年 1 月 31 日-2015 年 8 月 31 日に報道された移籍に関する記事が，移籍先チームの候補である，アーセナル，レアル・マドリード，マンチェスター・ユナイテッドのチームごとにまとめられている．それらの記事群を用いて，2015 年 7 月 1 日-2015 年 8 月 31 日の移籍期間終了時にリオネル・メッシ選手がどのチームに所属しているのかを予測する．予測の対象には，元々リオネル・メッシ選手が所属していたバルセロナに残留するという選択肢もある．

3.2 特徴量の抽出

特徴量の抽出にあたり，各記事データに対しては $a_{s,o}^{(t)}$ のように，時刻 t，選手 s，移籍先チーム o が適切に付加されていると仮定する．また，$a_{s,o}^{(t)}$ は記事情報として，記事の内容のテキストデータと，報道時期の情報，報道元の情報を含む．よって，例えば図 1 ようなデータ構造になっている．メッシ選手に関して，記事が移籍期間，移籍先チームごとに分類されていて，各記事が本文と，記事の公開日，掲載元 Web サイトのドメインを持つ．その構造が選手ごとに存在していることを表している．

全ての $a_{s,o}^{(t)}$ に対して，テキストデータは形態素解析とストップワードの除去を行なってから TF-IDF によってベクトル化する．

$$\mathrm{tf}_{a_{s,o},i}^{(t)} = \frac{a_{s,o}^{(t)}\text{における単語 } i \text{ の出現頻度}}{a_{s,o}^{(t)}\text{における全単語の出現頻度}}$$

df_i = 単語 i を含む文書数

また，$|A_{S,O}^{(T)}|$ を全ての報道期間の全ての主語と目的語を含む記事集合の数とする．これらを用いて，各単語における重み w_i は以下のようになる

$$w_i = \mathrm{tf}_{a_{s,o},i}^{(t)} \times \log \frac{|A_{S,O}^{(T)}|}{\mathrm{df}_i}$$

よって，$a_{s,o}^{(t)}$ のテキストデータは以下のように表される．

$$\mathbf{d}_{s,o}^{(t)} = (w_1, w_2, ..., w_n)$$

報道元に対応するベクトルは，One-hot-vector を用いる．つまり，記事 $a_{s,o}^{(t)}$ が i 番目の報道元から発表記事であった場合，次式のように全ドメインに対応するベクトルのう i 番目のの次元に対応する値のみが 1 となり，それ以外は 0 となる．

$$\mathbf{s}_{s,o}^{(t)} = (0, \ldots, 0, \overset{i}{\overset{\vee}{1}}, 0, \ldots, 0)$$

一般的にはドメイン数を m として次のようにかける．

$$\mathbf{s}_{s,o}^{(t)} = (s_1, s_2, ..., s_m), (s_k \in \{0, 1\}, \Sigma s_k = 1)$$

報道時期は，該当の移籍期間で，その選手に関する最初の記事が出た時刻 t_{first} からの経過日数を使用する．

$$\mathbf{t}_{s,o}^{(t)} = t - t_{first}$$

これらの特徴量を連結したベクトルによって，記事 $a_{s,o}^{(t)}$ を表現する．

$$\mathbf{x}_{s,o}^{(t)} = (\mathbf{d}_{s,o}^{(t)}, \mathbf{s}_{s,o}^{(t)}, \mathbf{t}_{s,o}^{(t)}) \tag{1}$$

3.3 学習と出力

3.2 節で述べた記事のベクトル表現 $\mathbf{x}_{s,o}^{(t)}(1)$ を用いて，学習にはロジスティック回帰モデルを用いて，各選手および移籍チームに対して，移籍するかしないかの 2 クラス分類問題として学習を行う．

$$f(\mathbf{x}_{s,o}^{(t)}) = \frac{1}{1 + e^{-\mathbf{W}^T \mathbf{x}_{s,o}^{(t)}}} \tag{2}$$

選手やチームは区別なく扱って，全て共通のモデルによって学習を行う．

学習には 2 クラス分類問題としてのアプローチを取るのに対して，出力には異なるアプローチをとる．なぜなら，求めるべきものは 1 記事に対する出力でもロジスティック回帰により得

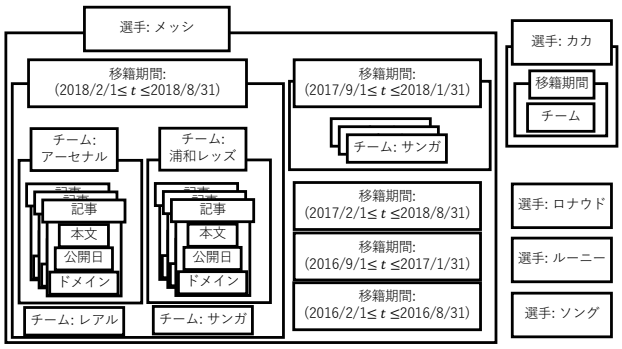

図 1　データの全体構造

られる確率値でもなく，その時期において得られる記事を統合した上で得られる，最も移籍が起きそうなチームであるからである．まず，特定の選手 s に対して，ある特定の時刻 t までの各チームの各記事について，一定の期間 $\tau(t-c \leqq \tau \leqq t)$ の全ての記事に対する出力を式 (2) のロジスティック回帰のモデルから得る．その上で，それらの出力結果を統合して出力チームを決定する．各記事データに対する出力結果を元に，チームごとのスコアを算出してそのスコアを基準に出力チームを決定する．特定期間におけるチーム o の記事集合 $A_{s,o}^{(\tau)}$ を用いて，スコア $F_{s,o}^{(t)}$ は $F(A_{s,o}^{(\tau)})$ と書くことができる．さらに以下のように，式 (1) を用いて各 $a_{s,o}^{(t_k)} \in A_{s,o}^{(\tau)}$ に対応する特徴ベクトル $\mathbf{x}_{s,o}^{(t_k)}$ を連結して $\mathbf{X}_{s,o}^{(t)}$ を定義し，$F_{s,o}^{(t)}$ を $F(\mathbf{X}_{s,o}^{(t)})$ と書くことができる．

$$\mathbf{X}_{s,o}^{(t)} = (\mathbf{x}_{s,o}^{(t_1)}, ..., \mathbf{x}_{s,o}^{(t_N)})$$

$$t_k \in \tau, (k = 1, ..., N)$$

$$t - c \leqq \tau \leqq t$$

ただし N は $A_{s,o}^{(\tau)}$ 内の記事数 $N_{s,o}^{(\tau)}$ を表す．

次に，スコア $F(X_{s,o}^{(t)})$ の説明を行う．複数の出力を統合する方法はいくつか考えることができるが，1 つは，以下のように各チームごとの出力の平均値をとる方法がある．

$$F_{mean}(\mathbf{X}_{s,o}^{(t)}) = \frac{\sum_{t_k \in \tau} f(\mathbf{x}_{s,o}^{(t_k)})}{N_{s,o}^{(t)}}$$

しかし，単純な平均であると，1 記事にしか出てこないチームで，かつその記事が高いスコアを獲得していた場合に，強くその影響を受けてしまう．そのようなケースは容易に考えられる．例えば，記事にたまたま 2 つのトピックが書かれていた場合に，1 つ目のトピックには，まさに対象の移籍に関する内容が書かれていて内容も移籍を強く示唆するものであり，記事に現れるチームに移籍する確率は高く，高いスコアを与えられるべきであるが，2 つ目のトピックは全くそれと関係のない内容で，記事に現れるチームは対象の移籍に全く関係ない．しかしながら，両方に大きなスコアがつけられてしまう．この時，後者のチームが，関連する記事集合内における別の記事中に現れる可能性は非常に低く，平均値をとってもこの記事のみがスコアに影響を与えることになる．結果として非常に高いスコアを得てしまう．複数の選手に関する移籍報道をまとめた記事などが多くあるため，これは想定すべきことである．このような事態を避けるために平均値以外の方法も用意しておくべきである．

偶然高いスコアを得た記事による影響を防ぐため，起こる可能性が高い移籍に対しての報道の数は比較的に多くなるという仮定に基づいて，記事数に応じてスコアに重みを加える関数を定義する．記事数に関する関数 $h(N)$ を導入して一般的に以下のように表すことができる．

$$F_{general}(\mathbf{X}_{s,o}^{(t)}) = h(N_{s,o}^{(t)}) \sum_{t_k \in \tau} f(\mathbf{x}_{s,o}^{(t_k)})$$

$h(N) = \frac{1}{N}$ の時，F_{mean} となる．また，$h(N) = 1$ の時には単純に各記事に対する出力値の和になる．これを F_{linear} とする．

$$F_{linear} = \sum_{t_k \in \tau} f(\mathbf{x}_{s,o}^{(t_k)})$$

スコアの平均値に対して線形な関数に対して，$h(N) = \log N$，とすると，より少ない記事数でも重要視するスコア F_{log} が得られ，$h(N) = \alpha N^2$ とするとより大きな記事数のみを重要視するスコア F_{quad} が得られる．

$$F_{log} = F_{linear} \times \log N$$

$$F_{quad} = F_{linear} \times \alpha N^2$$

その他にも、閾値を超える記事数をもつチームのみをスコアリングの対象とすることなども考えられる。これらのスコアリングによって複数記事を統合的に扱うことができる．スコアの計算にどの関数を選んだとしても，単純にスコアが最も高かったチームを出力チームとする．

また，ハイパーパラメータ θ を用意して，出力が閾値 θ を下回った場合は移籍先チームなしと出力する．また，s が t 時点でいたチーム $y_{s,0}^{(t)}$ を定義する．つまり，選手 s，時刻 t に対して以下のような出力 $\hat{y_s^{(t)}}$ を得る．

$$\hat{y_s^{(t)}} = \begin{cases} \arg\max_{o \in O} F(\mathbf{X}_{s,o}^{(t)}) & (\max_o F(\mathbf{X}_{s,o}^{(t)}) \geqq \theta) \\ y_{s,0}^{(t)} & (otherwise) \end{cases}$$

なぜこのような多クラス分類のタスクにおいて，2 クラス分類の学習方法をとっているのかについて補足しておく．今回のタスクでは，どのチームに移籍するかの多クラス分類問題として学習することが自然である．しかし，1 記事に出てくるチーム数はせいぜい数チームであり，このデータから 1000 チーム以上ある全てのチームに対するクラス分類を学習するのは難しい．出力を 1 記事だけで決定するよりも，1 記事からはその記事の表現と移籍の実現性の関係のみを学習する方が現実的である．そのため，先に述べたように，あらかじめ記事に移籍チームの候補を対応づけさせておくことで，共通モデルを用いた 2 クラス分類問題として学習することにした．

3.4 提案損失関数

通常ロジスティック回帰を用いた学習であるならば，損失関数は本研究のような場合は交差エントロピーを使用することが多い．だが，今回の学習を考えると，単純にどのデータに対しても同様の最適化を行うことは効果的ではない．なぜなら報道初期の情報などは，判断材料に乏しく最終的な移籍結果を必ずしも予測できているとは限らない．また，初期の報道には必ずしも正確性が求められているわけではない．また，飛ばし記事の多い報道機関は知られており，確信度は報道元によって顕著に見られるということも考えられる．よって，どの記事に対しても均質に正解ラベルに応じた最適化を行うのは得策ではない．そこで，その記事がどの程度確信があって報道されたものなのかを推定して，その確信度に応じて正解ラベルに対する最適化の度合いを調整する損失関数を以下に提案する．

$$L = \gamma(1 - g(\mathbf{z}')) + g(\mathbf{z}')(y - f(\mathbf{z}))^2$$

$\gamma \in [0, 1]$ ペナルティ係数

$\mathbf{z}$, $\mathbf{z}'$ が入力ベクトルの部分集合で y が正解ラベルである．

$g(\mathbf{z}')$ がその記事の確信度を [0, 1] で学習して，$f(\mathbf{z})$ が移籍確率を学習する．これは，あらかじめ入力ベクトルの部分集合 $\mathbf{z}'$ から確信度に類するものが学習できると仮定して入力ベクトルの要素から 2 つのベクトル $\mathbf{z}'$，$\mathbf{z}$ の要素をそれぞれ適当に選択して学習することを表している．損失関数の第 2 項目の $(y - f(\mathbf{z}))^2$ の部分は正解との二乗誤差である．確信度 $g(\mathbf{z}')$ が 1 に近ければ，つまり確信度が非常に高ければ，その二乗誤差がそのまま損失として扱われる．逆に，確信度が 0 に近い，つまり確信度が非常に低い場合は，二乗誤差の項は小さくなって，代わりにペナルティ係数 γ に比例する損失が加わることになる．ペナルティ係数が低ければ，二乗誤差の損失よりもペナルティ係数による損失の方が小さくなるため，全体的に確信度を非常に低く学習するようになる．逆に，ペナルティ係数が高ければ，二乗誤差の損失よりも確信度が小さいことのペナルティが大きくなり，確信度を高く学習するようになると期待される．二乗誤差と確信度が $[0, 1]$ の値をとる以上，第二項も $[0, 1]$ の値をとるため，ペナルティ係数の大きさは必ず $[0, 1]$ の間になくてはならない．これによって，確信度が低い場合は出力が正確であるかどうかは学習に重視されず，確信度が高ければ，誤差の影響が直接的に学習されることになる．直感的には，確信度に応じた損失を被る代わりに，各時点において曖昧な回答を許すことになる．

ペナルティ係数 γ について，以下詳細に議論する．出力 $(y - f(\mathbf{z}))^2 = \alpha$，$g(\mathbf{z}') = g$ とおくと損失 L は簡単に以下のように表せる．

$$\begin{aligned} L &= \gamma \times (1 - g(\mathbf{z}')) + g(\mathbf{z}') \times (y - f(\mathbf{z}))^2 \\ &= \gamma(1 - g) + \alpha g \end{aligned}$$

もし f が正確な値を取れば $\alpha = 0$ であるので，$g = 1$ となるように学習されることが望ましい．逆に f が不正確ならば，$\alpha > 0$ となり，この時，γ と α の大小関係によって，下のように g の望ましい振る舞いは変化する．

$$g = \begin{cases} 0 & (\gamma < \alpha) \\ 1 & (\alpha < \gamma) \end{cases}$$

つまり，γ が決定しているのは，α との大小関係により，g を 0 か 1 のどちらに近づけるようにするか，と言える．ここでもし γ が 0.25 より大きかった場合，$\alpha = 0.25$，つまり $f(z) = 0.5$ の時に確信度 $g = 1$ にするのがよいと学習されてしまう可能性がある．$f = 0.5$ は完全にランダムな出力であり，最も望ましくない学習である．よって γ は 0.25 より小さい値である必要がある．

したがって，この損失関数を使用することで，確信度が低い状況で無理に出力の値を正解に合わせて学習する必要がなくなる．そのため，そもそも正解か否かの正確な判断が非常に困難とされる記事を無理やり学習してしまって，本来学習すべきでない特徴を学習してしまうことを防ぐことができる．例えば，報道初期の記事に移籍決定と書いてあっても，それは飛ばし記事の可能性が高く，出版側も飛ばし記事という認識で書いている可能性がある．その記事に対して無理に学習をするのは，飛ばし記事である時点で正解不正解に関わらず，間違った学習になる．このような正当に学習すべきでない記事を、確信度を導入することで無理に学習しないでおくことが可能になる．また，各データから明示的に確信度が計算できることで結果の解釈可能性にも大きく貢献する．本研究においては，確信度が報道元と報道時期から推定できると仮定し，$\mathbf{z}'$ にそれらから抽出される特徴量を用い，$\mathbf{z}$ には報道記事のテキストデータから得られる特徴量を用いることにする．つまり，式 (1) の入力ベクトル $\mathbf{x}_{s,o}^{(t)} = (\mathbf{d}_{s,o}^{(t)}, \mathbf{s}_{s,o}^{(t)}, \mathbf{t}_{s,o}^{(t)})$ を用いて，以下のように設定する．

$$\mathbf{z}_{s,o}^{(t)} = (\mathbf{d}_{s,o}^{(t)})$$

$$\mathbf{z}_{s,o}^{(t)\prime} = (\mathbf{s}_{s,o}^{(t)}, \mathbf{t}_{s,o}^{(t)})$$

4 実　　験

4.1 デ ー タ

本研究のデータセット及び，サッカー選手の移籍の特徴について述べる．本研究においては，移籍が起こったことを確認する対象の Web 上の知識ベースを transfermarket (https://www.transfermarkt.com/) 上に記録された移籍の集合とした．データセット作成時には，このサイトからクローリングしてきた現在までの 15 年程度の累積した移籍データを元にして，各記事に対して該当の移籍が起こったかどうかの判定を行なった．移籍記事については，あらかじめ 20 のドメインを指定して，Google の検索 API を用いて 50 万記事程度の記事を収集した．クエリに選手の名前を用いることで，特定の選手に対応する記事を収集したため、この段階で各記事に 1 選手が対応している．今回の実験においては，便宜的に報道元という概念をこれらのドメインに対応させることとした．また，各記事に対して，記事の発表日抽出を行い，これを報道時期として利用する．

本研究においては，選手，移籍先チーム，報道元，報道時期ごとにデータを利用する必要がある．そのため，各記事に対して移籍先チームを確定させる必要がある．各記事がどのチームに移籍すると主張しているのかを正確に判断するのは困難であるため，今回は，記事内に現れたチームを全て，その記事が主張する移籍先チームだと判断することにした．例えば，記事内に 3 つのチーム名が現れていた場合は，その 1 つの記事は別々の移籍先チームに対する 3 つの別の記事として扱う．記事によっては，非常に多くのチームが現れるものもあってノイズになるので，記事中に現れるチーム数が 5 以下のものに制限してデータを作成した．記事内のチームを抽出する方法には Web 上の Wikification [10] サービスを用いた．Wikification とは，文書内の単語を実世界のエンティティに対応させてその実体を特定するエンティティリンキングの 1 種であり，各エンティティ

表 1　データ数と移籍数

データ		移籍	
データ総数	814000	移籍自体の総数	10276
移籍した	14000		903
移籍せず	800000		9373

に Wikipedia(https://en.wikipedia.org/) の各記事を対応させることでエンティティを特定するものが Wikification である．Wikification の誤差で望んでいないものが抽出される可能性もあるが，もともと設定した知識ベースにあるチームの集合のみを抽出対象とした．これによって各記事に対して，文書内にどのサッカーチームが現れているかを特定することができる．

次に，該当の移籍が起こったかどうかのラベルづけについて説明する．サッカー選手の移籍は，移籍期間が年間で 2 回，毎年同じ時期に固定されて設定されている．実際には，移籍期間は冬の移籍期間（1 月 1 日から 1 月 31 日）と夏の移籍期間（6 月 9 日から 9 月 1 日）に設定されている．例えば，4 月 30 日に公開された記事は，その年の夏の移籍期間の移籍に対応しており，10 月 3 日に公開された記事は次の年の冬の移籍期間の移籍に対応した記事であると言える．そのため，選手と報道時期が分かれば，その記事に対応した移籍が 1 つに定まり，対応した移籍でその選手がどのチームに移籍したのか，または移籍しなかったのかが分かる．これらの情報を元にして各記事に対して正解データのラベルづけを行うことができる．

また，学習の際には，学習データセットに工夫を加える．ここまでに述べた方法に従うと、1 つの記事に対して、記事内に現れるチーム数分に対応するデータが作成される．例えば、1 つの記事の中に，バルセロナ，アーセナルという 2 種類のサッカーチームに対応する表現が含まれているとする．その記事は移籍先の対象チームがバルセロナとされたデータと、アーセナルとされたデータの，同様の記事内容を持った 2 つのデータになる．しかし，学習データについては，正解チームがある場合は，それ以外のチームのデータは学習に使用しないことにした．つまり，上の例だと，正解移籍先のチームがバルセロナの場合，アーセナルのデータは削除する．この方が正解データを正しく学習できると判断した．なぜなら，この処理を行わない場合は正解である記事と全く同一の内容が，同時に不正解であると学習されてしまうからである．なお，記事に移籍先チームが存在しない場合は処理を行わない．ただし，検証時には同様のデータ処理を行なってはならない．そもそも検証時の時点では立場的にどの記事が正解なのかは分かっていないため，削除を行うのは不自然であるし，検証データが正解データに偏ってしまう．

データセットに関する注意として，検証データを適切に分割する必要がある．検証データは，検証の出力過程において，選手が各移籍期間単位でどのチームに移籍するのかを決定するため，移籍期間ごとにデータが全てまとまっていなくてはならない．

4.2　データ選別

表 1 に 4.1 節で作成したデータセットのデータ数とそこに含まれる移籍数を示す．

表 2　各段階のデータ数および割合

	フィルタリング後	アンダーサンプリング後
移籍総数	3992	815
移籍した	482	482
移籍せず	3510	333

表 3　最終的なデータ数

学習データ	88797	検証データ	25943
移籍した	10272	移籍した	2314
移籍せず	78525	移籍せず	23629

ここから，実際に使用したデータは，以下のような条件を満たすもののみを抽出して，選手，移籍期間ごとにグループ分けを行なった：(1) 5 チームより多くのチームが記事に現れない．(2) ある選手に対して，1 つの移籍期間に 50 データ以上のデータが存在する．(3) 移籍日時より発表日時が前である．上の 3 つの条件でフィルタリングを行なった結果，表 2 のような結果を得た．

次に，学習を適切に行わせるために、学習データにおける正のデータと負のデータ，すなわち，移籍が起こったデータと起こらなかったデータの割合の差を小さくするように調節を行う．表 2 に示すフィルタリング後のデータをそのまま使用すると，負のデータの割合が圧倒的に多いために，正のデータの学習がうまく行われない．一般にこのような性質のデータのことを不均衡データ [3, 4] と呼び，いくつかの有効な手法および，アルゴリズムが提案されている．

今回は，負のデータをアンダーサンプリング [9] を行なって数を減らす方法を採用した．これによって，データの不均衡性を取り除くことができる．サンプリング方法は，単純に，移籍が実際には起こらなかった移籍をランダムに 10% 程度選んでそれをデータとして使用して残りは使用しないようにした．4.1 節末に述べた理由で，サンプリングの対象はデータ単位ではなく，移籍単位で行う．ある移籍に対する記事は全て学習に使われるのが自然である．

最後に，前述のとおり，学習効率を上げるために，学習データに対してのみ，正例データを含む記事に対して，同じ記事における負例データを削除する処理を行う．学習データの方が正解ラベルのものの割合が高くなっていることが見て取れる．その結果得られたデータの分布を表 3 に示す．なお，サンプリング以降の数字はどの移籍がサンプリングされるかによって変化しうるのであくまで一例である．

4.3　実験設定

評価指標として，真陽性，真陰性，偽陽性，偽陰性を拡張した以下のものを定義する：(1) どこにも移籍しないと予測したが，どこかのチームに移籍した．(False Negative:FN) (2) どこにも移籍しないと予測して，どこのチームにも移籍しなかった．(True Negative:TN) (3) あるチームに移籍すると予測して，どこのチームにも移籍しなかった．(False Positive:FP) (4) あるチームに移籍すると予測して，そのチームに移籍した．(True Positive:TP) (5) あるチームに移籍すると予測して，別のチームに移籍した．(False Another Positive:FAP) またこの定義を

表 4 実験結果

	mean accuracy	mean TN	mean FN	mean TP	mean FP	mean FAP
Baseline	0.423 (0.034)	46.43 (4.77)	38.20 (5.80)	24.30 (4.02)	23.60 (3.80)	34.30 (5.01)
Add Feature	0.432 (0.032)	47.77 (5.28)	39.22 (5.06)	24.35 (3.80)	22.26 (3.84)	33.23 (4.69)
Loss Function	0.442 (0.028)	48.78 (5.49)	38.97 (6.24)	25.00 (3.76)	21.25 (4.34)	32.83 (5.20)

表 5 ハイパーパラメータ

ハイパーパラメータ	
スコアリング関数	F_{linear}
出力閾値 θ	0.3
提案損失関数　出力閾値 θ	0.2
ペナルティ係数 γ	0.2

使用して精度 (Accuracy) は以下のようにかける．

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FAP} + \text{FP} + \text{FN} + \text{TN}}$$

実験ではこれらの値を比較して考察を行う．

本研究では，まず，ベースライン手法として，記事のテキストデータのみを用いてロジスティック回帰で学習して実験を行う．次に，従来からのテキストデータベースの学習に対して，提案する特徴ベクトルを加えることで，学習の精度がどのように変化するのかを調査する．本研究においては，いくつかハイパーパラメータが存在するが，3.4 節での γ に関する議論に加えてチューニングを行い，表 5 のように決定した．

結果の評価について考えておくと，ベースラインと，2 つの特養量を付与した手法との比較は，その評価指標を単純に比較することができ，その特徴量が予測に貢献したのかどうかを判断すればよい．だが，損失関数を提案手法に変化させた時の比較をどのように行うのかについては少し考える必要がある．今回は，上の 2 つと出来るだけ同じ条件で比較したいので，確信度の影響が直接スコアに反映されるような形をとった．移籍確率 f に確信度 g の値をかけてそれを出力とすることとした．このようにして確信度の学習が予測に寄与しているのかどうかを確認できると考えた．よって今回の場合は，損失関数のにおける f と g にはロジスティック回帰を用いるので，各データ $\mathbf{x}_{s,o}^{(t)}$ に対するスコア $l(\mathbf{x}_{s,o}^{(t)})$ は以下のようにかける．

$$\begin{aligned} l(\mathbf{x}_{s,o}^{(t)}) &= f(\mathbf{x}_{s,o}^{(t)}) \times g(\mathbf{z}_{s,o}^{(t)\prime}) \\ &= \frac{1}{1+e^{-\mathbf{W}^T\mathbf{z}_{s,o}^{(t)}}} \times \frac{1}{1+e^{-\mathbf{W}^T\mathbf{z}_{s,o}^{(t)\prime}}} \\ &\quad (\text{ただし } \mathbf{x}_{s,o}^{(t)} = (\mathbf{z}_{s,o}^{(t)}, \mathbf{z}_{s,o}^{(t)\prime})) \end{aligned}$$

この方法であれば，上の 2 つの手法と同様に比較を行うことができる．また，損失関数の収束可能性についても結果を用いて確認する．

4.4 実験結果

実験では，先の設定に基づき，訓練・検証データの分割，訓練データのアンダーサンプリングを無作為に 100 回行なって，それぞれにおける各評価指標を計測し，その平均値を算出した．

実験結果は表 4 の通りである．また，括弧内は各評価指標における 100 個の標準偏差を表す．

以下，本結果から得られる知見と考察を述べる．まずは，予測タスクの達成について評価したい．すなわち，どの程度良くサッカー選手の移籍予測が的中しているのかを考える．本実験において使用した，各移籍における平均の移籍先候補のチーム数，つまり，予測対象の各 1 移籍に対する記事に現れたチームの数は平均で 20.5 チームであった．移籍しなかったケースを無視すると、無作為にチームを選択した場合の正解確率の期待値は 0.05 程度である．表からわかる通り，予測タスクにおける正答率は最も高いモデルで 0.44 程度なので，これを大きく上回っている．極端な例として、移籍しなかったケースの割合が大きいため、全てを移籍しなかったと出力すれば一定の精度が出ると考えられる．だが、表より TrueTeam も相当数存在するため、そのような極端な学習でないことがわかる．よって，少なくとも予測タスクをある程度達成することができたと言える．

続いて，それぞれの手法による比較を行う．結果から同一のデータで条件を揃えて実験を行うと，ベースライン手法，提案特徴量の付与，提案損失関数の使用の順に予測精度がわずかに上がっていることが見て取れる．すなわち，報道元ベクトルの情報と報道時期の情報が予測精度の向上に寄与する可能性があることが示唆された．加えて，True Team の値も同様の順番になっていることから，移籍が起こった場合についても効果があったことが分かる．

ここでは，提案損失関数についての結果を観察する．確信度 g の最終的な学習状態における出力結果は，全ケースで $[0.4904, 0.5445]$ の小さな範囲に収まっており，全てのデータが同程度の確信度に学習されていることが分かる．そのため，相対的なスコアの大小には影響は小さく，損失関数を変更しても結果に大きな差は出なかったのだと考えられる．また，学習後のモデルのパラメータで報道時期に対応するものを確認すると，その分布は $[-0.0327, 0.0322]$ となっており，標準化されたデータに対して掛け合わせるとほとんど 0 に近いため線型結合されたスコアに対して、大きな影響はでない．加えて，値が負である場合に関して，本来は時間が経つにつれて記事の確信度が増していくのが直感的だが，それとは逆の結果が現れている事になる．報道元に対応するパラメータも同様に小さい値を取っていた．また，最も平均値が高かったものに対応する報道元は，一般的に飛ばし記事が多いという評価を受けることが多い，イギリスの大衆紙「サン」の Web サイトであった．よって報道時期，報道元ともに，数値的にも意味的にも、想定していたように正しく確信度が学習できたとは言えない．

このようになった原因としては，確信度を計算するモデルが単純でありすぎたこと，そもそも，データセット内の記事ではどの時期にもどの報道元からもあまり偏りなく正解報道，不正解報道が行われていたことなどが考えられる．報道記事に関し

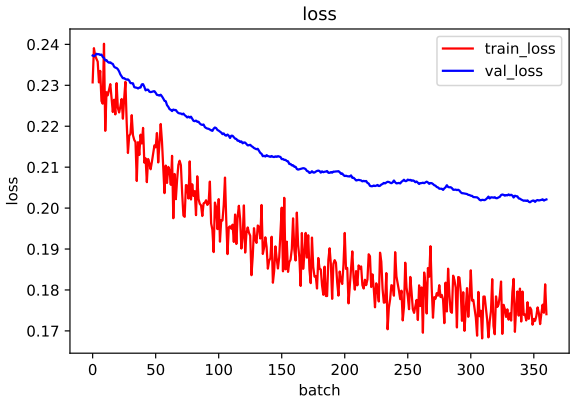

図 2　提案損失関数の損失推移

ては，収集には Web 上のハブ的なサイトも使用していたため，記事自体が 1 次ソースでないこともあったこと，サイトドメインを単純に報道元としたことが原因となったかもしれない．また，移籍と関係のない記事がノイズとなった可能性もある．

とはいえ，影響は小さくとも g の値は $[0.4904, 0.5445]$ の一定の幅を持って学習されていたので，提案損失関数にしてトータルでの精度が上がったことに対して，確信度が寄与している可能性はある．

次に，提案損失関数の収束可能性について考える．図 2 は，提案損失関数使用時の各バッチにおける損失の大きさのグラフである．本研究の学習では，学習モデルのキャパシティの低さに対して，入力ベクトルの次元数がデータ数に対して非常に多いため，収束は早く，すぐに過学習してしまう可能性が非常に高い．よって，エポックごとの損失ではなく，バッチごとの損失を可視化する．図から，学習が進むごとに検証データにおける誤差（val loss）が訓練データにおける誤差（train loss）と同様に減少して，傾きが 0 に近づいていっていることが見て取れる．よって，グラフから実験的に提案損失関数の収束性について確認することができた．

5　ま　と　め

本研究では，Web 上の記事を用いて将来における知識ベースの変化を予測する問題に取り組んだ．

特に，記事の報道元と報道時期が，記事の内容から予測される結果の確信度に大きな影響を与えると仮定し，この仮定を明示的に表現する損失関数および特徴量を提案した．

実験では，「サッカー選手の移籍予測」を題材として，記事の報道元と報道時期が予測精度に与える影響，および，提案する損失関数によって結果の確信度が正確に予測できるかを評価した．その結果，単純な方法で一定の精度で予測をすることができることが示せた．加えて，新しい特徴量を加えることで予測精度が向上することを示した．しかし，その特徴量から記事の正確な確信度を学習できたという確証を得ることはできなかった．

今後の課題としては，本研究において明確に確認できなかった記事の確信度が学習できることを示すため，今回は比較的単純に済ませた確信度を得るための特徴量の抽出をさらに工夫し複雑化すること，確信度を学習するためにロジスティック回帰以外の最適なモデルの選択を行うことが挙げられる．また，本研究の仮説が，より一般的な記事に対しても成立することを示すことも今後の課題である．

謝辞　本研究は JSPS 科研費 18H03244 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[4] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

[5] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

[6] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pages 1–4, 2012.

[7] O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. 1999.

[8] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9), 2011.

[9] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[10] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.

[11] G. Mishne, N. S. Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 155–158, 2006.

[12] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM, 2013.

[13] M. Samadi, P. P. Talukdar, M. M. Veloso, and M. Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *AAAI*, pages 222–228, 2016.

[14] G. Xiang, Z. Zheng, M. Wen, J. I. Hong, C. P. Rosé, and C. Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *ICWSM*, 2012.